Extracting Diagnosis from Japanese Radiological Report

Takeshi Imai[1], Yuzo Onogi PhD[2]

[1]The Graduate School of Interdisciplinary Information Studies, the University of Tokyo, Japan, [2]Clinical Bioinformatics Unit, Graduate School of Medicine, the University of Tokyo, Japan

Abstract

This study is aimed at extracting diagnosis with positive or negative assertion from radiological report written in Japanese Natural Language. We get frequency of verb patterns that indicate pos/neg assertion, and extract a rule in order of the occurrence.

We made customized dictionary of 36,152 terms relating to disease names or radiological findings, and tried to extract pairs of (pos/neg , disease and verb pattern) by using rules according to the most frequent pattern from 1,524/5,000 CT reports (each report consists of 15.1 words on the average). We tried only a few rules so far, and continue to find other rules.

Introduction

Automated extraction of diagnosis or findings from radiological report is expected to be useful not only for data-mining applications but also for clinicians to get summaries without oversights. But lexical analysis is very difficult especially for Japanese, because Japanese is an example of an agglutinative language, which is formed by concatenation of words, where each word is not separated by delimiter which is seen in inflected language such as English. So we just started to establish an efficient method instead of lexical analysis. In this study we introduce the first attempt to analyze the pattern for pos/neg assertion by using noun, verb, and parts of speech that indicate "denial".

Methods

We experimented with 5,000 CT reports (each report consists of 29.2 characters on average) from Department of Radiology, University of Tokyo Hospital.

We used Japanese tokenizer JUMAN, which is a program that segment Japanese into 'words' and tags these words with a part of speech.

We customized the dictionaries used in JUMAN by following resources as Japanese medical vocabulary

1) 'Standard Disease Names in Japan corresponding to ICD-10 for electronic medical record' published by MEDIS-DC.

2) 'Thesaurus for Medical and Health related Terms version 5' published in 2003 by JAMAS.

After JUMAN's processing, we got frequency of verb pattern that indicate pos/neg assertion then identified following rules from actual sentences fell in the most frequent pattern.

Rule 1) If each sentence is terminated with disease name, then recognize it as a positive assertion (usually, terminated with noun means it is important, in Japanese).

Rule 2) If each sentence is terminated with negative verb patterns, then search nearest disease name in the previous position, and regard it as a negative assertion.

We extracted 1,468 positive and 72 negative assertions in this method.

References

1. Y.Liu, Y.Satomura. Building a Controlled Health Vocabulary in Japanese.Method Inform Med 4,2001

2. M.Tanaka. A Mapping Approach of Japanese Medical Vocabulary to UMLS. 20th JCMI, 2000

3. Hripcsak G, Kuperman GJ, Friedman C. Extracting findings from narrative reports: software transferability and sources of physician disagreement. Methods Inf Med. 1998 Jan;37.